DOCUMENT RESUME

ED 391 812 | TM 024 151

| | |
|---|---|
| AUTHOR | Way, Walter D.; McKinley, Robert L. |
| TITLE | Development of Procedures for Resolving Irregularities in the Administration of the Listening Comprehension Section of the TOEFL Test. |
| INSTITUTION | Educational Testing Service, Princeton, N.J. |
| REPORT NO | ETS-RR-91-6; TOEFL-TR-3 |
| PUB DATE | Feb 91 |
| NOTE | 41p. |
| PUB TYPE | Reports - Evaluative/Feasibility (142) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | *Analysis of Covariance; *Bayesian Statistics; *English (Second Language); *Language Proficiency; Language Tests; Listening Comprehension; *Listening Comprehension Tests; Simulation; Test Construction; *Testing Problems; Test Interpretation |
| IDENTIFIERS | *Test of English as a Foreign Language |


ABSTRACT

Two procedures were developed to determine whether examinees in a given test center were affected by a testing irregularity on the Listening Comprehension section of the Test of English as a Foreign Language (TOEFL). One approach employed analysis of covariance (ANCOVA) on Listening Comprehension (Section 1) means using scores on Structure and Written Expression (Section 2) and scores on Reading and Vocabulary (Section 3) as covariates. The second procedure entailed a Bayesian approach that used prior information about performance at the center in question. Analyses using these two procedures were carried out with simulated data and data from actual testing irregularities at 5 centers involving 639 affected examinees and numerous comparisons. Results indicated that both ANCOVA and Bayesian approaches provided useful information. They usually agreed, but differences occurred in specific situations as discussed. Although both procedures should be incorporated into the operational procedures for resolving testing irregularities, results should be interpreted with caution, particularly if they produce discrepant results. An appendix contains five tables of summary statistics. (Contains 1 figure, 10 tables, and 3 references.) (Author/SLD)

# TOEFL

February 1991

# Technical Report

TR-3



## Development of Procedures for Resolving Irregularities in the Administration of the Listening Comprehension Section of the TOEFL Test

By Walter D. Way and Robert L. McKinley

Development of Procedures for Resolving Irregularities in the Administration of the Listening Comprehension Section of the TOEFL Test

Walter D. Way

Robert L. McKinley

Educational Testing Service

Princeton, NJ 08541

RR-91-6

## Abstract

The purpose of this study was to develop and evaluate two procedures to be used in determining whether examinees in a given test center are affected by a testing irregularity on the Listening Comprehension section of the Test of English as a Foreign Language (TOEFL). One approach employs analysis of covariance (ANCOVA) on Listening Comprehension (Section 1) means using scores on Structure and Written Expression (Section 2) and scores on Reading and Vocabulary (Section 3) as covariates. The second procedure entails a Bayesian approach that uses prior information collected about performance at the center in question. Analyses using these two procedures were carried out using both simulated data and data from actual testing irregularities.

The results of this study support the following conclusions. First, both the ANCOVA and Bayesian procedures appear to provide useful information related to the effects of testing irregularities on Section 1 of the TOEFL test; therefore, both procedures should be incorporated into the operational procedures for resolving testing irregularities. Second, the two procedures will usually agree about the effects of testing irregularities, although differences between the procedures may occur in situations where the Bayesian procedure indicates an effect of one scaled score point and the ANCOVA procedure does not indicate a statistically significant difference between the irregularity and comparison groups at some specified level. Finally, results based on the two procedures should be interpreted with caution, particularly in situations where the two procedures produce discrepant results.

The Test of English as a Foreign Language (TOEFL) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of more than thirty organizations, public and private, that were concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program, and in 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the GRE Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. Currently the Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. Local research may sometimes require access to TOEFL data. In such cases, the program may provide the data following approval by the Research Committee. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1990-91) members of the TOEFL Research Committee are:

| | |
|---|---|
| Patricia L. Carrell (Chair) | University of Akron |
| James Dean Brown | University of Hawaii |
| Patricia Dunkel | Pennsylvania State University |
| Fred Genesee | McGill University |
| Elliott Judd | University of Illinois at Chicago |
| Elizabeth C. Traugott | Stanford University |

TABLE OF CONTENTS

LIST OF TABLES

Table
No. Page



LIST OF FIGURES

Figure
No.

## Background

A successful administration of the Test of English as a Foreign Language (TOEFL) is one in which the difficulties that arise with any and all aspects of the testing process are successfully resolved. One area of the TOEFL test where problems often occur is in the administration of the Listening Comprehension section. Each administration, examinees and/or supervisors from several test centers complain about the quality of the Lis ening Comprehension recording. The problem may be due to a poor quality tape, inadequate tape player equipment, poor acoustics, or noisy disturbances outside the testing room. On average, approximately six Supervisor's Irregularity Reports (SIRs) that describe such problems with the Listening Comprehension section are received by the TOEFL Services Office after each administration and forwarded to the ETS Statistical Analysis area for review, analysis, and resolution. Because any given SIR can affect as many as several hundred examinees, and because there is an expense to the TOEFL program for retesting affected examinees when a testing irregularity is determined to have had an adverse impact, it is important that the statistical procedures applied to data from irregularities on the Listening Comprehension section be as valid as possible for the decision at hand.

### Current Resolution Procedures

Depending on the nature of the reported problem, Statistical Analysis staff employ one of several procedures for determining whether scores should be reported to examinees without any adjustment, whether scores should be adjusted, or whether examinees should be offered a retest. Although some of the reported Listening Comprehension problems involve only a small set of items, the majority of cases received involve problems with the entire Listening Comprehension section. When the performance on individual items is in question, an examination is made of the performance on the individual items and the performance on the rest of the section for both the potentially affected examinees at the center and the unaffected examinees who are either from the same test center or from other test centers. The result of this examination is a determination of whether the potentially affected examinees at the test center in question were adversely affected on the items in question.

For those cases in which the problems involve the entire Listening Comprehension section or a reasonably large subset of Listen Comprehension items, the procedure currently used to determine whether examinees were adversely affected is based simply on an examination of the mean and standard deviation of raw scores on the three sections. A comparison is made of the raw score means and standard deviations of the examinees in question on the three sections of TOEFL to the raw score means and standard deviations of the examinees at other centers. In cases where only a subset of the examinees at a center are potentially affected, a comparison is made of the raw score means and standard deviations of the affected and unaffected examinees at the same center.

If the Listening Comprehension raw score mean for the potentially affected group is, for argument's sake, 5 points lower than the Listening Comprehension raw score mean for an unaffected group at the same center, and the potentially affected group performed better on the other two sections of the TOEFL test, adverse impact is evident. However, such a clear-cut case rarely occurs in practice.

More often, the Listening Comprehension scores for the potentially affected examinees are fewer than 5 points lower than those for the comparison group, while the mean scores for Sections 2 and 3 are about the same. Thus, in many cases a simple comparison of section raw score means may not be sensitive enough to detect real difficulties in the administration of the Listening Comprehension section when the whole section is affected.

Also, in many cases it is necessary to use examinees at other centers, perhaps even in different countries, as the unaffected comparison group. Section score patterns vary across regions and language groups, which increases the risk of interpreting a legitimate score pattern as evidence of adverse impact. Because straightforward comparisons may be ineffective for identifying Listening Comprehension problems, it would seem that more statistically sophisticated procedures are warranted.

Because of the unique nature of the problem, there appears to be nothing in the literature that directly addresses the question of how irregularities in the Listening Comprehension section are to be resolved. The purpose of this study was to develop and evaluate two procedures to be used to determine whether examinees in a given test center are affected by a testing irregularity. One approach employs analysis of covariance (ANCOVA) on Listening Comprehension (Section 1) means using scores on Structure and Written Expression (Section 2) and scores on Reading and Vocabulary (Section 3) as covariates. The second procedure entails a Bayesian approach that uses prior information collected about performance at the center in question.

## Method

The two methods that were employed are ANCOVA using two covariates and a Bayesian procedure. The procedures were evaluated using Monte Carlo methods and with real data.

### ANCOVA Method

Instead of comparing the means of the three TOEFL sections for the potentially affected and unaffected groups, Section 2 and Section 3 scores are used as covariates in an ANCOVA of the Listening Comprehension raw score means for the two groups. If the potentially affected group consisted of all examinees at the test center, the group of unaffected examinees would consist of examinees at other centers in the same geographical region. If the group of potentially affected examinees consisted of a subset of examinees at a given test center, the group of unaffected examinees would consist of other examinees at the same center provided that the sample sizes were adequate.

In the ANCOVA procedure, scores on the Listening Comprehension section are regressed onto Section 2 and 3 scores separately for the potentially affected and the unaffected groups. If the regression coefficients associated with each group are homogeneous, then a pooled within-class regression coefficient is computed for each covariate (Winer, 1971). The multiple regression equation obtained for each group has the form:

$$Y'_{ij} = b_{xj}\,(X_{ij} - \bar{X}_j) + b_{zj}\,(Z_{ij} - \bar{Z}_j) + \bar{Y}_j\ , \qquad (1)$$

where $b_{xj}$ = the regression coefficient for group j for Section 2;

$b_{zj}$ = the regression coefficient for group j for Section 3;

$X_{ij}$ = Section 2 score for examinee i from group j;

$\bar{X}_j$ = mean Section 2 score for group j examinees;

$Z_{ij}$ = Section 3 score for examinee i from group j;

$\bar{Z}_j$ = mean Section 3 score for group j examinees;

$\bar{Y}_j$ = mean Listening Comprehension score for group j examinees; and

$Y'_{ij}$ = predicted score for examinee i from group j.

If $b_{x1}$ and $b_{x2}$ are homogeneous and $b_{z1}$ and $b_{z2}$ are homogeneous, then a multiple regression equation that includes the pooled within-class regression coefficients has the form:

$$Y'_{ij} = b_x (X_{ij} - X_j) + b_z (Z_{ij} - Z_j) + Y_j . \qquad (2)$$

The regression coefficients $b_x$ and $b_z$ are computed from the pooled within-class variances and covariances.

Assuming the within-class regression coefficients are homogeneous, an ANCOVA is performed by combining the variation and covariation due to group (treatment) and error and by computing a multiple regression equation on the combined data. The equation is of the form

$$Y'_{ij} = b'_x (X_{ij} - \bar{X}) + b'_z (Z_{ij} - \bar{Z}) + \bar{Y} . \qquad (3)$$

Then, the variation of residuals is computed about the above equation. The reduced variation due to group is obtained by subtracting the adjusted error variation from the variation of the residuals in the regression equation based on the combined data (Winer, 1971).

If it is found that the adjusted means for the Listening Comprehension scores are significantly different, and if the direction of the difference favors the unaffected group, it is inferred that the potentially affected group was affected. If the within-class regression coefficients are not homogeneous and pooled within-class regression coefficients cannot be compute , it appears as if the relationship between Listening Comprehension scores and Section 2 and/or Section 3 scores is different for the potentially affected and unaffected groups. If the coefficients turn out to be not homogeneous, the assumptions of the ANCOVA model are violated and the procedure cannot be considered valid, although it may still be instructive to examine the results of the test for the equality of the adjusted means.

Bayesian Procedure[1]

The ANCOVA method allows one to assess differences in mean scores on the Listening Comprehension section while controlling for differences in Section 2 and 3 scores. Although this method holds promise, it does not take into account the expected performance of examinees at the center and/or region in question. The proposed Bayesian procedure allows the incorporation of historical data at the center and/or region in question and generates probabilities that particular magnitudes of score effects (i.e., effect sizes) have occurred, given the historical information and the observed data.

---

1 For a detailed introduction to Bayesian methods in educational and psychological research, see Novick and Jackson (1974).

Univariate Case. Given a set of scores on Section 1, Bayes theorem states that

$$P(E_j|\underline{X}) = P(\underline{X}|E_j)P(E_j)/P(\underline{X}) \ , \tag{4}$$

where E is the size of the effect of an irregularity;

$\underline{X}$ is the vector of observed scores for the candidates in question;

$P(E_j|\underline{X})$ is the probability of effect size $E_j$ given observed scores $\underline{X}$;

$P(\underline{X}|E_j)$ is the probability of observing scores $\underline{X}$ given effect size $E_j$;

$P(E_j)$ is our prior belief as the probability of occurrence of effect size $E_j$; and

$P(\underline{X})$ is the probability of observing scores $\underline{X}$ regardless of size.

In such an application, $P(\underline{X})$ is given by

$$P(\underline{X}) = \sum_{j=1}^{n} P(\underline{X}|E_j)P(E_j) \ , \tag{5}$$

where n is the number of effect sizes.

Note that in the above equation E is discrete, rather than continuous. It would be possible to treat E as continuous, specify a distribution for E, and replace the summation in Equation 5 with integration. However, for the purposes of this study, we will treat E as discrete.

Evaluation of Equation 4 over a set of $E_j$ yields a set of corresponding probabilities. At that point, two procedures can be explored. One is to obtain a Bayes modal estimate of effect size by identifying the effect size with the highest corresponding probability. The alternative is to obtain an expected a posteriori (EAP) estimate of effect size by computing the expected value of E using

$$EAP = \sum_{j=1}^{n} E_j P(E_j|\underline{X}) \ . \tag{6}$$

Both procedures will be explored in this study.

At this point, the computation of one term in Equation 4, $P(\underline{X}|E_j)$, has not been discussed. In order to compute this term, it is necessary to specify a population distribution for raw score X. For the purposes of this study, we will assume X is distributed according to the normal distribution, with mean $\mu$ and standard deviation $\sigma$ set equal to the historical mean and standard deviation for a test center. $P(\underline{X}|E_j)$ is then computed using

$$P(\underline{X}|E_j) = \frac{1}{(2\pi)^{1/2}\sigma} \prod_{i=1}^{N} \exp[-0.5(X_i - \mu + E_j)^2/ \sigma^2] , \qquad (7)$$

where $X_i$ is the observed score for candidate i and

N is the number of affected candidates.

Multivariate Case. In order to incorporate information from examinee performance on TOEFL Sections 2 and 3, Equation 7 was extended so that the probability of observing scores on Section 1 ($\underline{X}_1$) was conditioned upon effect size $E_j$ as well as scores on Section 2 ($\underline{X}_2$) and Section 3 ($\underline{X}_3$). It was assumed that $X_1$, $X_2$, and $X_3$ are distributed multivariate normal, with means $\mu_1$, $\mu_2$, and $\mu_3$, and variance-covariance matrix $\underline{\Sigma}$. Equation 7 is then rewritten as

$$P(\underline{X}_1|E_j,\underline{X}_2,\underline{X}_2) = (2\pi)^{-3/2}|\underline{\Sigma}|^{-1/2} \prod_{i=1}^{N} \exp[\underline{x}_i'\underline{\Sigma}^{-1}\underline{x}_i] , \qquad (8)$$

where $\underline{x}_i' = [X_{1i} - \mu_1 + E_j, X_{2i} - \mu_2, X_{3i} - \mu_3]$, and

$$\underline{\Sigma} = \begin{bmatrix} \sigma^2{}_1 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma^2{}_2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma^2{}_3 \end{bmatrix} .$$

The Bayesian procedure employed for the present study incorporates information from performance on Sections 2 and 3 in calculating both Bayes modal and EAP estimates of effect size. This is accomplished by substituting $P(\underline{X}_1|E_j,\underline{X}_2,\underline{X}_3)$ obtained from Equation 8 for $P(\underline{X}|E_j)$ in Equation 5, which yields $P(\underline{X}_1|\underline{X}_2,\underline{X}_3)$ rather than $P(\underline{X})$. These quantities are then used with Bayes theorem to obtain the effect size posterior probability, $P(E_j|\underline{X}_1,\underline{X}_2,\underline{X}_3)$.

It should be pointed out that it would be possible to incorporate additional information into the Bayesian procedure, such as loss functions or information related to the variability of effect sizes. However, these potential features of the Bayesian procedure were not explored in this study.

Generation of Simulated Data

The study was conducted using Monte Carlo procedures. Simulated data sets consisting of scaled scores on the three sections were generated for five hypothetical test centers using trivariate normal distributions. Scaled scores rather than raw scores were generated because the available historical information for TOEFL scores is on the scaled score metric. Note, however, that the ANCOVA procedure could be applied to either raw scores (which may be available before converted scores are obtained) or to scaled scores. For two of the test centers (Centers A and B), scores on all three sections were based on historical means and standard deviations from actual test centers. The covariances were obtained from the correlations that appeared in the TOEFL test analysis report for the September 1988 administration. Scores on the Listening Comprehension section (Section 1) were generated to be unaffected by any irregularity in administration. These first two centers were used as the comparison test centers for the ANCOVAs. Scores for a third test center (Center C) were also generated to be unaffected by any irregularity in administration. For this test center, mean scores for the three TOEFL sections were generated to be equal to the means of the Center A and B generating means. Standard deviations for Center C were generated to be equal to the mean of the Center A and B generating standard deviations. The covariances between sections used to simulate the data for Center C were based on the same section intercorrelations as those used for the Center A and Center B simulations.

Scores for a fourth test center (Center D) were generated using the same means and standard deviations as for Center C for Sections 2 and 3. However, for this center, scores for the Listening Comprehension section were generated to simulate a center that was affected by an irregularity. The mean Listening Comprehension raw score for this center was therefore set to be lower than the mean for Listening Comprehension for Center C. The covariances between sections were the same as for Centers A, B and C. For a fifth hypothetical test center (Center E), scores for Sections 2 and 3 were again generated using the means and standard deviations used with Center C, while the generating mean for the Listening Comprehension section was the same as for Center D. However, the generating covariances for Center E were based on correlations between Sections 1 and 2, and correlations between Sections 1 and 3 that were lower than those used in generating the data for the other test centers.

The data were simulated using three different sets of sample sizes. For the first set, a total of 200 scores on Sections 1, 2, and 3 were simulated for Centers A and B, while 100 scores were simulated for each of Centers C, D, and E. For the second set, 100 scores were simulated for Centers A and B, while 50 scores were simulated for each of Centers C, D, and E. For the third

set, 50 scores were simulated for Centers A and B, while 25 scores were simulated for Centers C, D, and E. Within each set of sample sizes, five replications were carried out. For Centers A, B, and C, the same Section 1 generating means were used for all five replications. For Centers D and E, the mean Section 1 scores were simulated to range from one scaled score point lower in replication 1 (effect size = -1) to five scaled score points lower in replication 5 (effect size = -5).

The generating parameters and obtained summary statistics for the simulated data are displayed in Tables A.1 to A.5 of the appendix. A total of 75 data sets (five test centers X three sample sizes X five replications) were generated for the study.

Analysis of the Simulated Data

Of the five hypothetical test centers, Centers A and B always served as the comparison group of unaffected examinees for the ANCOVA procedure. Centers C, D, and E served as the SIR groups and were evaluated using both ANCOVA and Bayesian procedures. Center C was evaluated as the unaffected center to determine if the procedures would lead to the conclusion that the center was unaffected. Centers D and E were evaluated as affected centers to determine if the procedures would lead to the conclusion that the centers were affected. The ANCOVA and Bayesian procedures were evaluated independently.

Data from Real Irregularities

The ANCOVA and Bayesian procedures were also applied to data from five actual TOEFL irregularities. All irregularities occurred in foreign test centers in administrations between December 1988 and May 1989. Descriptions of the irregularities for each of the selected test centers are given below.

Real Test Center 1 (REAL1). According to the supervisor for one of the testing rooms at this center, it was necessary to replace a defective tape recorder and to repeat questions 1-10 of the Listening Comprehension section. A total of 57 examinees were affected. For the ANCOVA procedure, 275 examinees in other rooms at Center REAL1 were used as the comparison group. For the Bayesian procedure, historical means and standard deviations were based on 644 examinees who were tested at that center between January 1987 and October 1988.

Real Test Center 2 (REAL2). The tape recording for the Listening Comprehension section was reported as unclear during several questions in one of the rooms at this test center. A total of 29 candidates were affected. For the ANCOVA procedure, the rest of the center (72 candidates) was used as the comparison group. For the Bayesian procedure, historical means and standard deviations were based on the scores of 274 candidates who were administered the TOEFL test at Center REAL2 between January 1987 and November 1988.

Real Test Center 3 (REAL3). A total of 348 examinees in two testing rooms complained about the quality of the tape recording for the Listening Comprehension test. The comparison group for the ANCOVA consisted of the other 733 candidates at the center. For the Bayesian procedure, historical means and standard deviations were based on the scores of 908 candidates who were administered the TOEFL test at Center REAL3 between May 1987 and March 1989.

Real Test Center 4 (REAL4). Several examinees testing at this center complained about the quality of the tape for Section 1. A total of 123 examinees at Center REAL4 were affected. The comparison group for the ANCOVA consisted of scores for a total of 266 examinees testing at other centers throughout the same country. For the Bayesian procedure, historical means and standard deviations were based on the scores of 233 candidates who were administered the TOEFL test at Centers in the same city as Center REAL4 between March 1987 and April 1989.

Real Test Center 5 (REAL5). The testing supervisor reported poor quality of the Section 1 tape recording. A total of 82 examinees at the center were affected. The comparison group for the ANCOVA consisted of scores for a total of 239 examinees testing at other centers throughout the same country. For the Bayesian procedure, historical means and standard deviations were based on the scores of 203 candidates who were administered the TOEFL test at centers in the same city as Center REAL5 between March 1987 and April 1989.

Table 1 summarizes the information related to these irregularities, including the scaled score means, standard deviations, and sample sizes of the affected and comparison groups, as well as the historical means and standard deviations of the test centers where the irregularities occurred.

Limitations

It should be noted that both the simulated and real data sets investigated in this study may not generalize completely to the variety of situations that occur in real TOEFL administrations. Irregularities that occur with the Listening Comprehension section of TOEFL are often difficult to resolve in a straightforward manner regardless of the statistical method employed. However, it was expected that the data examined in this study would be sufficient to yield information about the utility of the ANCOVA and Bayesian procedures for many of the typical irregularities that occur with the TOEFL test.

Table 1
Observed Summary Statistics for the Irregularity Groups, the Comparison Groups, and the Historical Information for the Real Data SIRs

| Irregularity Group | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Section 1 | | Section 2 | | Section 3 | | Section Corrs. | | |
| Test Center | N | X | SD | X | SD | X | SD | $r_{12}$ | $r_{13}$ | $r_{23}$ |
| REAL1 | 57 | 52.86 | 6.98 | 54.21 | 6.31 | 51.84 | 5.66 | .79 | .65 | .73 |
| REAL2 | 29 | 56.72 | 5.97 | 54.69 | 6.64 | 51.45 | 6.92 | .87 | .89 | .90 |
| REAL3 | 348 | 50.95 | 6.62 | 55.52 | 6.44 | 54.59 | 5.80 | .72 | .70 | .80 |
| REAL4 | 123 | 52.41 | 7.10 | 53.95 | 6.65 | 54.09 | 5.85 | .66 | .58 | .70 |
| REAL5 | 82 | 53.76 | 7.33 | 55.74 | 7.75 | 54.29 | 6.72 | .70 | .76 | .79 |
| **Comparison Group** | | | | | | | | | | |
| | | Section 1 | | Section 2 | | Section 3 | | Section Corrs. | | |
| Test Center | N | X | SD | X | SD | X | SD | $r_{12}$ | $r_{13}$ | $r_{23}$ |
| REAL1 | 275 | 54.08 | 6.66 | 54.60 | 6.86 | 52.89 | 6.90 | .79 | .81 | .84 |
| REAL2 | 72 | 56.35 | 5.59 | 54.51 | 7.08 | 51.21 | 6.53 | .70 | .74 | .86 |
| REAL3 | 733 | 51.99 | 6.25 | 54.36 | 6.26 | 53.84 | 5.61 | .71 | .70 | .77 |
| REAL4 | 266 | 52.36 | 6.18 | 54.32 | 6.41 | 55.44 | 5.15 | .62 | .62 | .75 |
| REAL5 | 239 | 57.13 | 5.76 | 56.09 | 7.47 | 54.21 | 7.11 | .79 | .77 | .82 |
| **Historical Information** | | | | | | | | | | |
| | | Section 1 | | Section 2 | | Section 3 | | Section Corrs.[a] | | |
| Test Center | N | X | SD | X | SD | X | SD | $r_{12}$ | $r_{13}$ | $r_{23}$ |
| REAL1 | 644 | 55.00 | 7.42 | 55.38 | 8.08 | 54.08 | 6.99 | .71 | .68 | .81 |
| REAL2 | 274 | 57.51 | 5.86 | 55.23 | 7.07 | 53.36 | 6.61 | .71 | .68 | .81 |
| REAL3 | 908 | 53.25 | 6.46 | 54.57 | 6.60 | 55.94 | 5.91 | .71 | .68 | .81 |
| REAL4 | 233 | 53.48 | 7.22 | 53.78 | 7.65 | 55.16 | 6.25 | .71 | .68 | .81 |
| REAL5 | 203 | 57.42 | 6.12 | 54.91 | 7.31 | 53.91 | 7.97 | .71 | .68 | .81 |

[a]Historical section intercorrelations were assumed to be equal to those reported for foreign examinees in a recent TOEFL Test Analysis Report.

# Results

## Simulated Data - ANCOVA Procedure

The results of the ANCOVA procedure applied to the simulated data are summarized in Tables 2 to 4. Table 2 summarizes the results for SIR Center C (the unaffected center), Table 3 summarizes the results for SIR Center D (the affected center with lower section 1 means), and Table 4 summarizes the results for SIR Center E (the affected center with both lower section 1 means and lower section intercorrelations). In each of these tables, the comparison group consisted of data for Centers A and B combined.

Center C. From the left-hand columns of Table 2, it can be seen that for all data sets the hypothesis of homogeneous within-class regression coefficients was not rejected at a .10 level of significance. Thus, the tests for equality of slopes appeared to produce results that were consistent with the parameters used to generate the Center C data.

In the right-hand columns of Table 2 are the results of the tests for the equality of the adjusted means between Center C and Centers A and B combined. For combined sample sizes of 500 and 250, all tests resulted in accepting the null hypothesis at a .15 level of significance. For the combined sample size of 125, the hypothesis of equal adjusted means would have been rejected at a .05 level of significance for replication 2, replication 4, and replication 5. However, in Table A.3 of the appendix, the observed means for Center C in replications 4 and 5 indicate that the significant ANCOVA results in Table 2 for these replications were the result of higher, rather than lower, Section 1 adjusted means for Center C compared to the adjusted means for Centers A and B combined. Thus, for the combined sample sizes of 125, only in replication 2 was there evidence that an irregularity occurred for Center C.

Center D. On the basis of the procedures used to simulate the Center D data, it was not expected that differences in the within-group regression coefficients would be statistically significant. The results of the tests for the equality of slopes shown in Table 3 appear consistent with this expectation. These results indicate that only for replication 5 with the combined sample size of 250 was the hypothesis of homogeneous within-class coefficients rejected at a significance level of .05.

The results of the tests for the equality of the adjusted means between Center D and Centers A and B combined (Table 3) also appeared to be consistent with the procedures used to simulate the Center D data. For combined sample sizes of 500, the null hypothesis was rejected at a significance level of .05 for all replications. However, for combined sample sizes of 250 the null hypothesis was not rejected at a significance level of .05 for replication 1 (effect size = -1), and for combined sample sizes of 125 the null hypothesis was not rejected at a significance level of .05 for replications 1, 2, or 3 (effect sizes = -1, -2, and -3). Thus, it appeared that only when the sample sizes were 100 for the affected group and 400 for the comparison group was the

Table 2
Summary of ANCOVA Results for Test Center C versus
Test Centers A and B Combined at Various Sample Sizes

Sample Sizes = 400 (Centers A & B), 100 (Center C)

| | Test for Equality of Slopes (Degrees of freedom = 2, 494) | | | Test for Equality of Adj. Means (Degrees of freedom = 1, 494) | | |
|---|---|---|---|---|---|---|
| Repl. | Mean Sq. | F | Prob>F | Mean Sq. | F | Prob>F |
| 1 | 11.46 | 0.50 | 0.61 | 0.27 | 0.01 | 0.91 |
| 2 | 5.53 | 0.22 | 0.80 | 16.66 | 0.68 | 0.41 |
| 3 | 18.51 | 0.86 | 0.42 | 5.75 | 0.27 | 0.61 |
| 4 | 8.81 | 0.37 | 0.69 | 13.96 | 0.59 | 0.44 |
| 5 | 33.94 | 1.56 | 0.21 | 2.84 | 0.13 | 0.72 |

Sample Sizes = 200 (Centers A & B), 50 (Center C)

| | Test for Equality of Slopes (Degrees of freedom = 2, 244) | | | Test for Equality of Adj. Means (Degrees of freedom = 1, 244) | | |
|---|---|---|---|---|---|---|
| Repl. | Mean Sq. | F | Prob>F | Mean Sq. | F | Prob>F |
| 1 | 6.33 | 0.26 | 0.77 | 0.00 | 0.00 | 0.99 |
| 2 | 1.08 | 0.04 | 0.96 | 4.47 | 0.17 | 0.68 |
| 3 | 3.48 | 0.15 | 0.86 | 0.27 | 0.01 | 0.92 |
| 4 | 18.81 | 0.74 | 0.48 | 47.25 | 1.85 | 0.18 |
| 5 | 12.27 | 0.51 | 0.60 | 0.71 | 0.03 | 0.86 |

Sample Sizes = 100 (Centers A & B), 25 (Center C)

| | Test for Equality of Slopes (Degrees of freedom = 2, 119) | | | Test for Equality of Adj. Means (Degrees of freedom = 1, 119) | | |
|---|---|---|---|---|---|---|
| Repl. | Mean Sq. | F | Prob>F | Mean Sq. | F | Prob>F |
| 1 | 26.06 | 1.11 | 0.33 | 57.88 | 2.47 | 0.12 |
| 2 | 9.70 | 0.49 | 0.61 | 133.64 | 6.82 | 0.01 |
| 3 | 52.76 | 2.19 | 0.12 | 3.63 | 0.15 | 0.70 |
| 4 | 32.75 | 1.63 | 0.20 | 196.85 | 9.69 | 0.00 |
| 5 | 0.44 | 0.02 | 0.98 | 116.25 | 4.99 | 0.03 |

Table 3
Summary of ANCOVA Results for Test Center D versus
Test Centers A and B Combined at Various Sample Sizes

Sample Sizes = 400 (Centers A & B), 100 (Center D)

| | Test for Equality of Slopes (Degrees of freedom = 2, 494) | | | Test for Equality of Adj. Means (Degrees of freedom = 1, 494) | | |
|---|---|---|---|---|---|---|
| Repl. | Mean Sq. | F | Prob>F | Mean Sq. | F | Prob>F |
| 1 | 0.43 | 0.02 | 0.98 | 102.16 | 4.27 | 0.04 |
| 2 | 15.14 | 0.56 | 0.57 | 258.37 | 9.65 | 0.00 |
| 3 | 36.44 | 1.71 | 0.18 | 492.44 | 23.04 | 0.00 |
| 4 | 34.81 | 1.38 | 0.25 | 1130.96 | 44.77 | 0.00 |
| 5 | 33.53 | 1.48 | 0.23 | 1406.52 | 61.89 | 0.00 |

Sample Sizes = 200 (Centers A & B), 50 (Center D)

| | Test for Equality of Slopes (Degrees of freedom = 2, 244) | | | Test for Equality of Adj. Means (Degrees of freedom = 1, 244) | | |
|---|---|---|---|---|---|---|
| Repl. | Mean Sq. | F | Prob>F | Mean Sq. | F | Prob>F |
| 1 | 7.79 | 0.32 | 0.73 | 49.23 | 2.03 | 0.16 |
| 2 | 33.60 | 1.33 | 0.27 | 321.61 | 12.69 | 0.00 |
| 3 | 7.13 | 0.29 | 0.75 | 129.30 | 5.29 | 0.02 |
| 4 | 3.40 | 0.14 | 0.87 | 490.45 | 19.65 | 0.00 |
| 5 | 77.63 | 3.37 | 0.04 | 760.50 | 32.35 | 0.00 |

Sample Sizes = 100 (Centers A & B), 25 (Center D)

| | Test for Equality of Slopes (Degrees of freedom = 2, 119) | | | Test for Equality of Adj. Means (Degrees of freedom = 1, 119) | | |
|---|---|---|---|---|---|---|
| Repl. | Mean Sq. | F | Prob>F | Mean Sq. | F | Prob>F |
| 1 | 12.94 | 0.52 | 0.59 | 6.81 | 0.28 | 0.60 |
| 2 | 13.54 | 0.70 | 0.50 | 67.53 | 3.53 | 0.06 |
| 3 | 10.69 | 0.41 | 0.67 | 88.80 | 3.43 | 0.07 |
| 4 | 26.76 | 1.34 | 0.27 | 150.73 | 7.51 | 0.01 |
| 5 | 16.82 | 0.78 | 0.46 | 244.25 | 11.44 | 0.00 |

ANCOVA procedure powerful enough to detect a difference of one scaled score point in the affected group. For affected and comparison sample sizes of 50 and 200, respectively, the ANCOVA procedure detected differences due to effect sizes of two or more scaled score points at a .05 level of significance. For affected and comparison sample sizes of 25 and 100, the ANCOVA detected only differences due to effect sizes of -4 and -5 at a significance level of .05.

Center E. Center E was the only center where the data were simulated so that covariate slopes would differ from those of the comparison group (Centers A and B). In Table 4, it can be seen that differences between the generating covariate slopes were not necessarily reflected in the statistical analyses of the observed data. For combined sample sizes of 500, the test for equality of slopes was rejected in four of the five replications at a .05 level of significance. However, for combined sample sizes of 250 for only two of the five replications were the differences in the covariate slopes statistically significant at a .05 level of significance. For combined sample sizes of 125, the hypothesis of equal covariate slopes was rejected at a .05 level of significance for only one of the five replications. These results suggest that the application of the ANCOVA procedure in actual irregularities may not detect violations to the equality of covariate slopes, particularly when the combined sample size for the irregularity and comparison groups is less than 500. It should be noted that the generating correlations for Center E were .52 between Sections 1 and 2, and .54 between Sections 1 and 3. These correlations were .15 lower than the corresponding generating correlations for Centers A and B.

Although strictly speaking the tests for the equality of the adjusted means are uninterpretable for Center E, the patterns of the results in Table 4 are quite similar to those found for Center D in Table 3. For the combined sample sizes of 500 and combined sample sizes of 250, the hypothesis of equal adjusted means was rejected in replications 2, 3, 4, and 5. For combined sample sizes of 125, the adjusted means for Center E were found to be significantly different from the adjusted means for Centers A and B combined at a .05 level of significance in replications 3, 4, and 5. Note, however, that for 7 of the 11 cases where differences in the adjusted means were found to be statistically significant, the ANCOVA procedures could not be considered valid on the basis of the slopes test. In analyzing an actual irregularity, it would be tempting to draw conclusions if the test of the adjusted means was statistically significant even if the test was not strictly appropriate. The fact that the affected center's covariate slopes differed from those of the comparison center and the adjusted Section 1 means were different would be reasonably compelling evidence that an irregularity had taken place.

Table 4
Summary of ANCOVA Results for Test Center E versus
Test Centers A and B Combined at Various Sample Sizes

Sample Sizes = 400 (Centers A & B), 100 (Center E)

| | Test for Equality of Slopes (Degrees of freedom = 2, 494) | | | Test for Equality of Adj. Means (Degrees of freedom = 1, 494) | | |
|---|---|---|---|---|---|---|
| Repl. | Mean Sq. | F | Prob>F | Mean Sq. | F | Prob>F |
| 1 | 40.18 | 1.61 | 0.20 | 47.17 | 1.88 | 0.17 |
| 2 | 84.90 | 3.21 | 0.04 | 100.14 | 3.75 | 0.05 |
| 3 | 70.68 | 2.95 | 0.05 | 703.23 | 29.14 | 0.00 |
| 4 | 167.79 | 6.61 | 0.001 | 1049.79 | 40.42 | 0.00 |
| 5 | 107.21 | 4.59 | 0.01 | 2899.45 | 122.41 | 0.00 |

Sample Sizes = 200 (Centers A & B), 50 (Center E)

| | Test for Equality of Slopes (Degrees of freedom = 2, 244) | | | Test for Equality of Adj. Means (Degrees of freedom = 1, 244) | | |
|---|---|---|---|---|---|---|
| Repl. | Mean Sq. | F | Prob>F | Mean Sq. | F | Prob>F |
| 1 | 38.45 | 1.49 | 0.23 | 1.33 | 0.05 | 0.82 |
| 2 | 4.39 | 0.16 | 0.85 | 137.33 | 5.08 | 0.03 |
| 3 | 44.31 | 1.77 | 0.17 | 383.32 | 15.21 | 0.00 |
| 4 | 173.83 | 6.23 | 0.00 | 483.83 | 16.63 | 0.00 |
| 5 | 97.09 | 3.91 | 0.02 | 740.21 | 29.12 | 0.00 |

Sample Sizes = 100 (Centers A & B), 25 (Center E)

| | Test for Equality of Slopes (Degrees of freedom = 2, 119) | | | Test for Equality of Adj. Means (Degrees of freedom = 1, 119) | | |
|---|---|---|---|---|---|---|
| Repl. | Mean Sq. | F | Prob>F | Mean Sq. | F | Prob>F |
| 1 | 27.40 | 1.05 | 0.35 | 1.98 | 0.08 | 0.78 |
| 2 | 29.43 | 1.37 | 0.26 | 62.84 | 2.91 | 0.09 |
| 3 | 102.13 | 3.72 | 0.03 | 185.11 | 6.45 | 0.01 |
| 4 | 19.73 | 0.91 | 0.41 | 614.22 | 28.37 | 0.00 |
| 5 | 66.88 | 2.50 | 0.09 | 428.46 | 15.63 | 0.00 |

Simulated Data - Bayesian Procedure

Representation of prior information. To implement the Bayesian procedure, prior probabilities of effect sizes ranging from -8 to 8 were specified. Because the magnitudes of the prior probabilities specified for each effect size influence the results of the Bayesian procedure, three different sets of prior probabilities were explored. Table 5 displays the probabilities of different discrete effect sizes based on these three different priors: normal, uniform, and stacked. A graph of the prior probabilities based on these priors is given in Figure 1. For each of the three sets of priors shown in Table 5, analyses were carried out using 9 of the 45 simulated data sets for Centers C, D, and E. The results of the preliminary analyses are given in Table 6. These data suggest that the form of the prior makes little difference in the resulting EAP and Bayes modal estimates of effect size. For only two of the nine data sets investigated were the Bayes model estimates of effect size different depending upon the prior used: Center C in replication 1 (C/25/R1), and Center E in replication 2 (E/25/R2). For each of these data sets the sample size was 25 and the influence of the observed data compared to the prior was relatively dilute. On the basis of these preliminary analyses, the normal prior was chosen for all analyses because it was most consistent with the authors' a priori expectations of how effect sizes would be distributed in potential SIR centers.

Table 5
Probabilities of Different Discrete Effect Sizes
Under Three Different Prior Distributions

| Effect Size | Normal Prior | Uniform Prior | Stacked Prior |
|---|---|---|---|
| -8 | 0.0057 | 0.0588 | 0.0300 |
| -7 | 0.0088 | 0.0588 | 0.0300 |
| -6 | 0.0186 | 0.0588 | 0.0300 |
| -5 | 0.0332 | 0.0588 | 0.0300 |
| -4 | 0.0542 | 0.0588 | 0.0300 |
| -3 | 0.0823 | 0.0588 | 0.0300 |
| -2 | 0.1052 | 0.0588 | 0.0300 |
| -1 | 0.1245 | 0.0588 | 0.1500 |
| 0 | 0.1350 | 0.0588 | 0.2800 |
| 1 | 0.1245 | 0.0588 | 0.1500 |
| 2 | 0.1052 | 0.0588 | 0.0300 |
| 3 | 0.0823 | 0.0588 | 0.0300 |
| 4 | 0.0542 | 0.0588 | 0.0300 |
| 5 | 0.0332 | 0.0588 | 0.0300 |
| 6 | 0.0186 | 0.0588 | 0.0300 |
| 7 | 0.0088 | 0.0588 | 0.0300 |
| 8 | 0.0057 | 0.0588 | 0.0300 |

# Figure 1

## Effect Size Priors

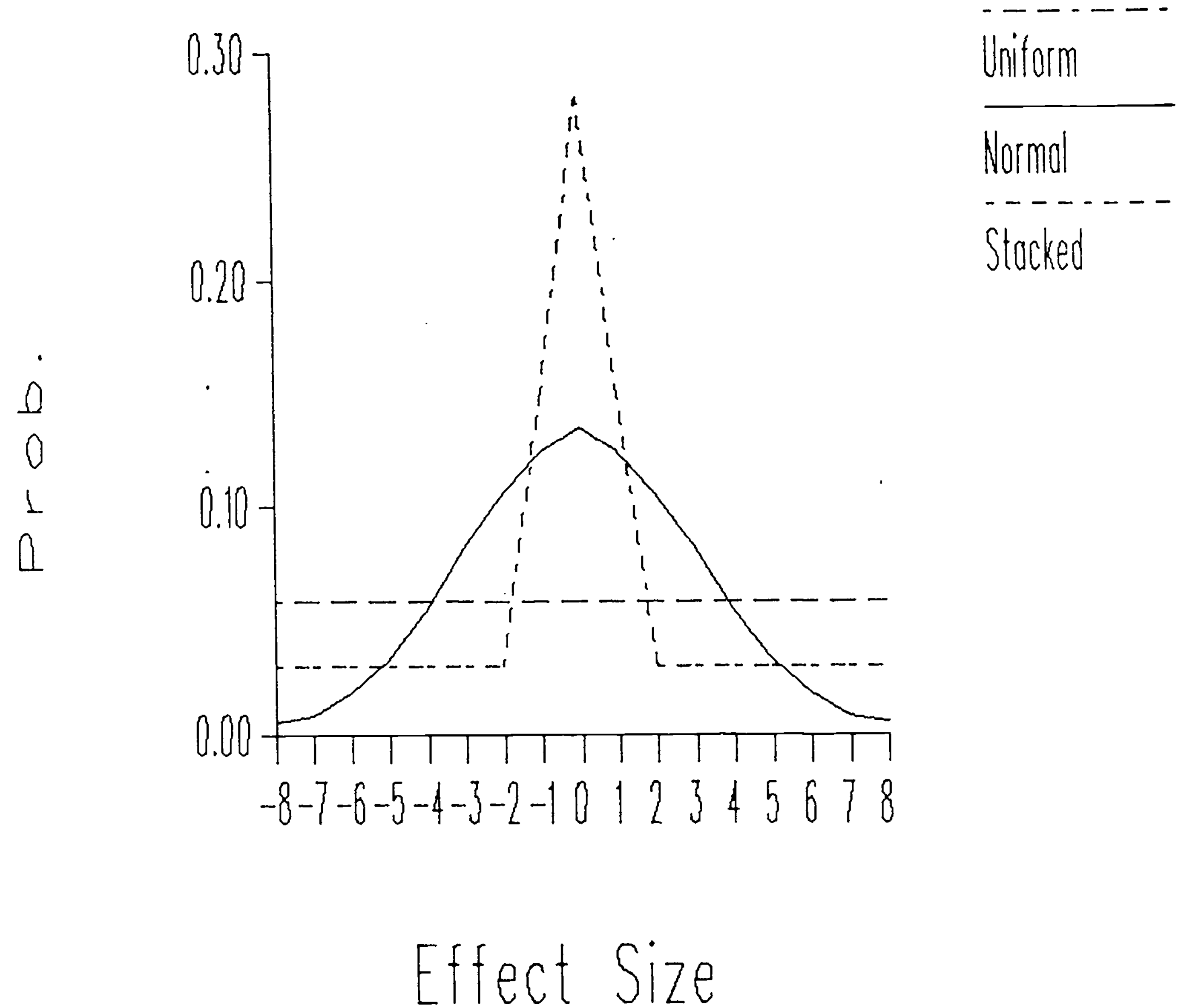

Table 6
Expected A Posteriori (EAP) and Bayes Modal Estimates
of Effect Sizes for Nine Different Simulated Data Sets

| Center / N / Repl. | Normal Prior EAP | Normal Prior Modal | Uniform Prior EAP | Uniform Prior Modal | Stacked Prior EAP | Stacked Prior Modal |
|---|---|---|---|---|---|---|
| C/100/R1 | 0.16 | 0.00 | 0.17 | 0.00 | 0.10 | 0.00 |
| D/100/R1 | 1.19 | 1.00 | 1.23 | 1.00 | 1.00 | 1.00 |
| E/100/R2 | 0.89 | 1.00 | 0.91 | 1.00 | 0.76 | 1.00 |
| C/50/R1 | 0.23 | 0.00 | 0.24 | 0.00 | 0.15 | 0.00 |
| D/50/R1 | 1.24 | 1.00 | 1.30 | 1.00 | 0.88 | 1.00 |
| E/50/R2 | 1.41 | 1.00 | 1.48 | 1.00 | 1.01 | 1.00 |
| C/25/R1 | -0.46 | 0.00 | -0.52 | -1.00 | -0.26 | 0.00 |
| D/25/R1 | 1.23 | 1.00 | 1.36 | 1.00 | 0.73 | 1.00 |
| E/25/R2 | 2.06 | 2.00 | 2.26 | 2.00 | 1.54 | 1.00 |

Center C. The results of applying the Bayesian procedure to the Center C data sets can be seen in the left-hand columns of Table 7. For sample sizes of 100, the posterior Bayes modal estimate of effect size was zero for all five replications. The EAP estimates ranged from -0.11 to 0.28. For sample sizes of 50, the Bayes modal estimates were zero for three replications and -1.00 for two replications. However, for sample sizes of 25, the Bayes modal estimate of effect size was 3.0 for replication 2 and -2.00 for replications 4 and 5. These results are consistent with the ANCOVA results for the same data sets, and suggest that when the number of examinees in the group affected by an irregularity is 25 or fewer, the results of either the Bayesian or ANCOVA procedure should be interpreted with caution.

Centers D and E. The results of the Bayes procedure for Centers D and E in Table 7 are quite consistent across replications. For these centers, perfect consistency with the simulation procedures would result in Bayes modal estimates increasing sequentially from one in replication 1 to five in replication 5. In general, the observed Bayes modal estimates of effect sizes closely approximated this pattern. In particular, for five of the six replication 1 data sets for Centers D and E, the Bayes modal estimate was consistent with the effect size used in simulating the data. In addition, for none of the Center D and E data sets was either the EAP or Bayes modal estimate of effect size below zero. Thus, the Bayesian procedure appeared to be very successful in identifying the irregularity effects in the simulated data.

Table 7
Expected A Posteriori (EAP) and Bayes Modal Estimates of Effect Sizes For Simulated Test Centers C, D, and E

Sample Size = 100

| | Center C | | Center D | | Center E | |
|---|---|---|---|---|---|---|
| | EAP | Modal | EAP | Modal | EAP | Modal |
| Repl.1 | 0.16 | 0.00 | 1.19 | 1.00 | 0.85 | 1.00 |
| Repl.2 | 0.19 | 0.00 | 1.59 | 2.00 | 0.89 | 1.00 |
| Repl.3 | 0.19 | 0.00 | 2.37 | 2.00 | 2.89 | 3.00 |
| Repl.4 | -0.11 | 0.00 | 3.94 | 4.00 | 3.78 | 4.00 |
| Repl.5 | 0.28 | 0.00 | 4.61 | 5.00 | 6.43 | 6.00 |

Sample Size = 50

| | Center C | | Center D | | Center E | |
|---|---|---|---|---|---|---|
| | EAP | Modal | EAP | Modal | EAP | Modal |
| Repl.1 | 0.23 | 0.00 | 1.24 | 1.00 | 0.37 | 0.00 |
| Repl.2 | -0.56 | -1.00 | 2.36 | 2.00 | 1.41 | 1.00 |
| Repl.3 | 0.07 | 0.00 | 1.86 | 2.00 | 3.07 | 3.00 |
| Repl.4 | -1.18 | -1.00 | 3.20 | 3.00 | 3.04 | 3.00 |
| Repl.5 | -0.32 | 0.00 | 4.10 | 4.00 | 4.09 | 4.00 |

Sample Size = 25

| | Center C | | Center D | | Center E | |
|---|---|---|---|---|---|---|
| | EAP | Modal | EAP | Modal | EAP | Modal |
| Repl.1 | -0.46 | 0.00 | 1.23 | 1.00 | 0.85 | 1.00 |
| Repl.2 | 2.92 | 3.00 | 2.08 | 2.00 | 2.06 | 2.00 |
| Repl.3 | -0.32 | 0.00 | 1.98 | 2.00 | 2.84 | 3.00 |
| Repl.4 | -2.29 | -2.00 | 3.29 | 3.00 | 5.66 | 6.00 |
| Repl.5 | -1.85 | -2.00 | 3.48 | 3.00 | 4.37 | 4.00 |

Comparison of the ANCOVA and Bayesian Procedures - Simulated Data

Based on the simulated data, it appeared that the Bayesian procedure was more successful than the ANCOVA procedure at correctly identifying the presence or absence of irregularity effects. The major difference between the two procedures occurred in the analyses for Centers D and E when the simulated effect size was a single scaled score point. However, this difference can be attributed to the statistical approaches taken by the two methods. In the ANCOVA procedure, a traditional hypothesis testing approach is taken. The tendency in this case is to use a traditional alpha level, such as .05 or .10. On the other hand, in the Bayesian procedure, the EAP and Bayes modal

estimates are estimates of the magnitude of effect that appear most likely on the basis of the data and the prior information about the test center. Neither of these statistics addresses whether a traditional null hypothesis ought to be rejected with a specified type I error rate. If one used the ANCOVA procedure with a less traditional alpha level, such as .20 or .30, the results of the procedure might agree better with results based on the Bayesian procedure. Similarly, in applying the Bayesian procedure, one could make a determination of whether the combined probabilities of effect sizes less than or equal to zero was sufficiently small based on a specified level of "significance." This would tend to produce results with the Bayesian procedure that were more in agreement with those based on the ANCOVA procedure.

Real Data - ANCOVA Procedure

The results of applying the ANCOVA procedure to the real data are given in Table 8. The results for each of the selected test centers are discussed in the paragraphs that follow.

Center REAL1. For this center, the test for the equality of the slopes was rejected at a .05 level of significance. Thus, the result of the test for equality of the adjusted means is questionable. However, there is not sufficient evidence on the basis of the ANCOVA procedure to conclude that examinees were affected by the SIR for Center REAL1.

Center REAL2. Neither the test for equality of slopes nor the test for equality of the adjusted means was statistically significant, as indicated in Table 8. Thus, it appears that the irregularity for Center REAL2 did not have an adverse affect on the examinees in question.

Center REAL3. For this center, there is strong evidence that the testing irregularity affected examinee scores on Section 1. While the slope parameter estimates for the SIR and comparison groups were not significantly different, the test of equality of the adjusted means was statistically significant at an alpha level of .001.

Center REAL4. From the data in Table 8, it appears that the irregularity in Center REAL4 did not have an effect on candidates' Section 1 scores. Neither the test for equality of slopes nor the test for equality of the adjusted means was significant at a .15 level of significance.

Center REAL5. In Table 8, it can be seen that both the tests for equality of slopes and equality of the adjusted means were statistically significant at $\alpha = .05$. Despite the questionableness of the ANCOVA procedure, there is a fairly strong indication that the testing irregularity did affect the scores obtained by the candidates at Center REAL5.

Table 8
Summary of ANCOVA Results for Irregularity Test Centers
Versus Comparison Test Centers

| SIR Center | Test for Equality of Slopes | | | | Test for Equality of Adj. Means | | | |
|---|---|---|---|---|---|---|---|---|
| | df | Mean Sq. | F | Prob>F | df | Mean Sq. | F | Prob>F |
| REAL1 | (2,326) | 53.82 | 3.71 | 0.03 | (1,326) | 18.20 | 1.23 | 0.27 |
| REAL2 | (2,95) | 11.81 | 0.99 | 0.38 | (1,95) | 1.04 | 0.09 | 0.77 |
| REAL3 | (2,1075) | 7.90 | 0.44 | 0.64 | (1,1075) | 802.77 | 45.24 | 0.00 |
| REAL4 | (2,383) | 33.74 | 1.44 | 0.24 | (1,383) | 39.42 | 1.68 | 0.20 |
| REAL5 | (2,315) | 55.14 | 4.03 | 0.02 | (1,315) | 658.24 | 47.26 | 0.00 |

Real Data - Bayesian Procedure

The expected a posteriori (EAP) and Bayes modal estimates of effect sizes for the actual SIR test centers are given in Table 9. For Center REAL3 both estimates indicate an effect size of about two scaled score points. For Center REAL5 both estimates indicate an effect size of about four scaled score points. Thus, for these two test centers, the evidence from the Bayesian procedure strongly suggests that the irregularities had an effect on candidates Section 1 scores. However, the results of the Bayesian procedure for the other three test centers are less clear cut. The EAP estimates of effect size are between 0 and 1 for Centers REAL1 (0.87), REAL2 (0.57), and REAL4 (0.80). For Centers REAL1 and REAL4 the Bayes modal estimate of effect size is 1.00, while the Bayes modal estimate for Center REAL2 is 0. Thus, a strict interpretation of these data would suggest that a SIR effect did occur for Centers REAL1 and REAL4 and did not for Center REAL2.

Table 9
Expected A Posteriori (EAP) and Bayes Modal Estimates
of Effect Sizes For Actual SIR Test Centers

| Center | EAP | Modal |
|---|---|---|
| REAL1 | 0.87 | 1.00 |
| REAL2 | 0.57 | 0.00 |
| REAL3 | 2.01 | 2.00 |
| REAL4 | 0.80 | 1.00 |
| REAL5 | 3.99 | 4.00 |

Comparison of the ANCOVA and Bayesian Procedures - Real Data

As with the simulated data, the noteworthy difference between the ANCOVA and Bayesian procedures occurred in cases where the Bayes modal estimate of effect size was 1.00 and the results of the ANCOVA did not suggest that the

adjusted means for the SIR and comparison groups were significantly different. This discrepancy would seem to be primarily due to inherent differences in the approach of the two methods rather than to any discrepancies in how the data were analyzed using the two methods. Note, however, that the real data case differed from the simulated data case in one important respect. Whereas in the simulations the same means and standard deviations were used to generate data and to represent the historical information, with the real data there is no guarantee that historical means and standard deviations used with the Bayesian procedure will be similar to comparison group means and standard deviations in the ANCOVA procedure. For example, in Table 1 it can be seen that the historical Section 1 means are higher than the observed Section 1 means of the ANCOVA comparison group for every test center. To test whether these differences were of any significance, the Bayesian procedure was carried out using the comparison group data as if it were the SIR group data. Table 10 contains the resulting EAP and Bayes modal estimates. These data indicate that for Centers REAL4 and REAL5, effect sizes of 1.00 were detected even though no irregularity was reported for the comparison groups. One possible reason for this finding is that for both Centers REAL4 and REAL5, the samples for the comparison group data and the historical data consisted of data taken from several test centers located in the same city or country rather than data from the same test center. These data may not have been as dependable as the data for the other test centers used in the study.

Table 10
Expected A Posteriori (EAP) and Bayes Modal Estimates of Effect Sizes For Comparison Test Centers

| Center | EAP | Modal |
|---|---|---|
| REAL1 | 0.04 | 0.00 |
| REAL2 | 0.17 | 0.00 |
| REAL3 | 0.26 | 0.00 |
| REAL4 | 1.33 | 1.00 |
| REAL5 | 0.99 | 1.00 |

## Discussion

Overall, the results of this study strongly support the use of the two proposed procedures for resolving testing irregularities on Section 1 of the TOEFL test. The two procedures provided similar evidence when applied to both simulated and real data, and differences between the two procedures appeared primarily to be due to differences in the way in which an "effect" was statistically determined. A major advantage of having data from both procedures is that one can be used to confirm the results based on the other. In applications to actual testing irregularities, agreement of results based on both the ANCOVA procedure and the Bayesian procedure would serve as solid evidence for a particular decision about whether an "effect" had actually occurred. Furthermore, because the two procedures depend upon different data sources, there may be situations when the available data for one of the procedures will be more relevant for a particular irregularity than the available data for the other. However, for cases in which application of the two procedures results in conflicting evidence, a decision must still be made. In these cases, there are several considerations to keep in mind.

### Context of the Irregularity

In many cases, the context in which the irregularity occurred can be used as a guide for interpreting results of the ANCOVA or Bayesian procedure. For example, for Center REAL1 in this study, it was necessary to replace a defective tape recorder and replay questions 1-10 of Section 1. In this case, because the offending questions were repeated, it would seem unlikely that the irregularity had an adverse affect on candidate scores.

### Characteristics of the Samples Used

An important consideration in using either the Bayesian or the ANCOVA procedure is the characteristics of the comparative samples. For example, with the ANCOVA procedure, if an entire center is affected by an irregularity it may be necessary to consider a comparison group from one or more centers in the same city, country, or even geographical region. Similarly, for the Bayesian procedure, if there is not sufficient historical data about a particular test center, then it may be necessary to use historical data for centers that may or may not share the candidate characteristics (e.g., native language or native country) of the affected center. In these cases, the data used to compare the data for the SIR center may be less satisfactory, and results from analyses based on these data should be interpreted carefully. A similar caveat relates to sample size. The results of the simulations using the ANCOVA procedure suggested that when sample sizes were 25 for the irregularity group and 100 for the comparison group, simulated effect sizes of 1 and 2 were generally not detected (see Tables 3 and 4). While this was partly due to sampling error in the data simulation, in practice the level of significance used to evaluate the ANCOVA procedure could take into account the number of examinees in the irregularity and comparison groups. Note that this caution also applies to situations when the comparison group consists of large

numbers of examinees. If the comparison group consists of 10,000 examinees, then the ANCOVA procedure will be powerful but not necessarily valid, that is, a relatively trivial difference in scaled scores may be found to be statistically significant.[2]

Operational Implementation of the Procedures

In general, implementation of the ANCOVA procedure requires little more than a statistical computing package to carry out the needed calculations. In addition, it makes little difference whether the ANCOVA procedure is applied to examinee raw scores or converted scores. This is a distinct advantage since there is often a delay between when a timely resolution of the irregularity is needed and when examinee converted scores are received.

On the other hand, the application of the Bayesian procedure is limited to examinee converted scores, since the available historical information for TOEFL test centers is limited to scaled score means and standard deviations. For this reason, the current version of the computer program that implements the Bayesian procedure will need to be modified to convert raw scores to scaled scores using the conversion table for a given administration.

A final consideration in the operational implementation of the ANCOVA and Bayesian procedures concerns the designation of the comparison data. As previously mentioned, it is not always possible to obtain a group of valid comparison examinees from the same test center for the ANCOVA procedure, nor to obtain stable estimates of a particular test center's historical mean and standard deviation for the Bayesian procedure. In fact, for domestic test centers, this information is basically impossible to come by, since there is no reason to expect that the ethnic makeup of either a comparison group or a historical group of candidates will share the ethnic makeup of an irregularity group, even if the data come from the same test center. However, for irregularities occurring in foreign test centers, the identification of plausible "surrogate" centers to serve for comparisons using either the ANCOVA procedure or the Bayesian procedure must be done carefully, and could certainly use the guidance of additional research.

---

2 A potentially useful approach to the problem of power in the ANCOVA analyses, suggested by L. Stricker (personal communication, July 27, 1989), might be to estimate and consider the statistical power of the particular ANCOVA analysis (Cohen, 1977).

## Conclusions

The results of this study support the following conclusions. First, both the ANCOVA and Bayesian procedures appear to provide useful information related to the effects of testing irregularities on Section 1 of the TOEFL test; therefore, both procedures should be incorporated into the operational procedures for resolving testing irregularities. Second, the two procedures will usually agree about the effects of testing irregularities, although differences between the procedures may occur in situations where the Bayesian procedure indicates an effect of one scaled score point and the ANCOVA procedure does not indicate a statistically significant difference between the irregularity and comparison groups at some specified level. Finally, results based on the two procedures should be interpreted with caution, particularly in situations where the two procedures produce discrepant results.

## References

Cohen, J. (1977). Statistical Power Analysis for the Behavioral Sciences. (Rev. Ed.) New York: Academic Press.

Novick, M.R., & Jackson, P.H. (1974). Statistical Methods for Educational and Psychological Research. New York: McGraw-Hill Book Co.

Winer, B.J. (1971). Statistical Principles in Experimental Design. (2nd Ed.) New York: McGraw-Hill Book Co.

Appendix A

Summary Statistics for the Simulated Data Sets

Table A.1
Summary Statistics for the Simulated Data Sets
for Test Center A at Sample Sizes of 200, 100, and 50

Sample Size = 200

| Data Set | Section 1 | | Section 2 | | Section 3 | | Section Corrs. | | |
|---|---|---|---|---|---|---|---|---|---|
| | X | SD | X | SD | X | SD | $r_{12}$ | $r_{13}$ | $r_{23}$ |
| Parms. | 49.93 | 6.10 | 54.01 | 6.06 | 52.02 | 6.02 | .67 | .69 | .81 |
| Repl.1 | 49.75 | 6.01 | 53.90 | 6.08 | 52.18 | 6.14 | .68 | .69 | .84 |
| Repl.2 | 49.63 | 6.36 | 53.57 | 5.76 | 51.60 | 5.74 | .64 | .65 | .78 |
| Repl.3 | 50.24 | 6.26 | 54.64 | 6.13 | 52.45 | 5.93 | .70 | .70 | .80 |
| Repl.4 | 49.24 | 6.51 | 53.45 | 6.77 | 51.75 | 6.16 | .69 | .66 | .82 |
| Repl.5 | 50.27 | 6.10 | 54.74 | 6.45 | 52.85 | 5.80 | .71 | .72 | .84 |

Sample Size = 100

| Data Set | Section 1 | | Section 2 | | Section 3 | | Section Corrs. | | |
|---|---|---|---|---|---|---|---|---|---|
| | X | SD | X | SD | X | SD | $r_{12}$ | $r_{23}$ | $r_{13}$ |
| Parms. | 49.93 | 6.10 | 54.01 | 6.06 | 52.02 | 6.02 | .67 | .69 | .81 |
| Repl.1 | 50.83 | 6.62 | 54.23 | 6.94 | 53.03 | 7.15 | .76 | .69 | .84 |
| Repl.2 | 49.84 | 6.36 | 53.37 | 6.16 | 52.02 | 5.96 | .70 | .67 | .81 |
| Repl.3 | 50.24 | 5.80 | 53.80 | 5.75 | 51.64 | 5.80 | .60 | .65 | .83 |
| Repl.4 | 50.29 | 6.90 | 54.76 | 6.49 | 53.04 | 6.38 | .75 | .78 | .84 |
| Repl.5 | 49.89 | 6.13 | 55.03 | 6.00 | 53.06 | 6.45 | .63 | .73 | .79 |

Sample Size = 50

| Data Set | Section 1 | | Section 2 | | Section 3 | | Section Corrs. | | |
|---|---|---|---|---|---|---|---|---|---|
| | X | SD | X | SD | X | SD | $r_{12}$ | $r_{23}$ | $r_{13}$ |
| Parms. | 49.93 | 6.10 | 54.01 | 6.06 | 52.02 | 6.02 | .67 | .69 | .81 |
| Repl.1 | 48.88 | 5.03 | 54.52 | 5.73 | 52.76 | 4.87 | .37 | .51 | .65 |
| Repl.2 | 49.44 | 5.16 | 53.38 | 6.00 | 52.24 | 6.32 | .64 | .65 | .78 |
| Repl.3 | 49.12 | 5.66 | 53.74 | 6.50 | 51.26 | 6.10 | .62 | .63 | .76 |
| Repl.4 | 49.66 | 5.65 | 52.84 | 5.95 | 51.06 | 6.59 | .62 | .64 | .76 |
| Repl.5 | 48.88 | 7.08 | 52.92 | 6.81 | 50.74 | 6.53 | .73 | .74 | .78 |

Table A.2
Summary Statistics for the Simulated Data Sets
for Test Center B at Sample Sizes of 200, 100, and 50

| Data Set | Section 1 X | Section 1 SD | Section 2 X | Section 2 SD | Section 3 X | Section 3 SD | $r_{12}$ | $r_{13}$ | $r_{23}$ |
|---|---|---|---|---|---|---|---|---|---|
| Sample Size = 200 | | | | | | | | | |
| Parms. | 45.28 | 7.55 | 46.45 | 9.40 | 47.08 | 8.52 | .67 | .69 | .81 |
| Repl.1 | 45.73 | 7.37 | 47.23 | 8.94 | 47.82 | 8.23 | .64 | .68 | .80 |
| Repl.2 | 45.96 | 7.96 | 46.78 | 9.49 | 47.41 | 8.32 | .67 | .66 | .82 |
| Repl.3 | 45.75 | 7.46 | 46.52 | 9.61 | 47.91 | 8.24 | .68 | .73 | .82 |
| Repl.4 | 44.92 | 7.30 | 46.15 | 9.26 | 46.58 | 8.84 | .63 | .69 | .80 |
| Repl.5 | 44.87 | 7.04 | 47.09 | 9.27 | 47.60 | 8.58 | .65 | .67 | .81 |

| Data Set | Section 1 X | Section 1 SD | Section 2 X | Section 2 SD | Section 3 X | Section 3 SD | $r_{12}$ | $r_{23}$ | $r_{13}$ |
|---|---|---|---|---|---|---|---|---|---|
| Sample Size = 100 | | | | | | | | | |
| Parms. | 45.28 | 7.55 | 46.45 | 9.40 | 47.08 | 8.52 | .67 | .69 | .81 |
| Repl.1 | 43.28 | 7.55 | 44.50 | 9.58 | 45.76 | 7.84 | .65 | .67 | .79 |
| Repl.2 | 44.30 | 8.07 | 44.27 | 8.47 | 45.31 | 8.36 | .68 | .70 | .81 |
| Repl.3 | 45.00 | 7.67 | 47.09 | 10.07 | 47.94 | 9.12 | .66 | .70 | .84 |
| Repl.4 | 45.01 | 7.67 | 44.91 | 9.96 | 46.01 | 8.72 | .65 | .65 | .88 |
| Repl.5 | 45.83 | 8.28 | 46.10 | 9.16 | 46.65 | 9.35 | .70 | .73 | .82 |

| Data Set | Section 1 X | Section 1 SD | Section 2 X | Section 2 SD | Section 3 X | Section 3 SD | $r_{12}$ | $r_{23}$ | $r_{13}$ |
|---|---|---|---|---|---|---|---|---|---|
| Sample Size = 50 | | | | | | | | | |
| Parms. | 45.28 | 7.55 | 46.45 | 9.40 | 47.08 | 8.52 | .67 | .69 | .81 |
| Repl.1 | 43.76 | 7.95 | 45.24 | 9.11 | 44.96 | 8.59 | .77 | .66 | .86 |
| Repl.2 | 44.40 | 7.88 | 45.78 | 9.36 | 46.70 | 7.95 | .73 | .80 | .80 |
| Repl.3 | 44.34 | 7.84 | 44.72 | 8.10 | 45.30 | 7.09 | .61 | .69 | .83 |
| Repl.4 | 44.22 | 8.27 | 47.42 | 10.49 | 48.30 | 9.06 | .80 | .80 | .88 |
| Repl.5 | 45.54 | 6.36 | 47.64 | 9.71 | 47.68 | 8.28 | .59 | .62 | .78 |

Table A.3
Summary Statistics for the Simulated Data Sets
for Test Center C at Sample Sizes of 100, 50, and 25

Sample Size = 100

| Data Set | Section 1 | | Section 2 | | Section 3 | | Section Corrs. | | |
|---|---|---|---|---|---|---|---|---|---|
| | X | SD | X | SD | X | SD | $r_{12}$ | $r_{13}$ | $r_{23}$ |
| Parms. | 47.61 | 6.83 | 50.23 | 7.73 | 49.55 | 7.27 | .67 | .69 | .81 |
| Repl.1 | 47.38 | 6.70 | 50.30 | 8.24 | 49.42 | 7.91 | .66 | .68 | .81 |
| Repl.2 | 47.29 | 6.91 | 50.18 | 8.80 | 49.37 | 7.52 | .80 | .81 | .85 |
| Repl.3 | 47.44 | 6.79 | 50.04 | 7.15 | 49.85 | 7.59 | .65 | .75 | .82 |
| Repl.4 | 47.76 | 6.33 | 50.67 | 7.80 | 49.26 | 7.01 | .66 | .72 | .80 |
| Repl.5 | 46.54 | 7.23 | 48.88 | 8.30 | 48.66 | 7.36 | .71 | .77 | .81 |

Sample Size = 50

| Data Set | Section 1 | | Section 2 | | Section 3 | | Section Corrs. | | |
|---|---|---|---|---|---|---|---|---|---|
| | X | SD | X | SD | X | SD | $r_{12}$ | $r_{23}$ | $r_{13}$ |
| Parms. | 47.61 | 6.83 | 50.23 | 7.73 | 49.55 | 7.27 | .67 | .69 | .81 |
| Repl.1 | 46.64 | 6.56 | 48.94 | 7.57 | 48.66 | 6.89 | .67 | .70 | .84 |
| Repl.2 | 47.58 | 7.33 | 49.40 | 7.98 | 48.58 | 7.26 | .69 | .69 | .87 |
| Repl.3 | 47.80 | 7.07 | 50.84 | 9.84 | 49.78 | 8.67 | .77 | .82 | .82 |
| Repl.4 | 48.92 | 7.32 | 50.22 | 7.60 | 49.72 | 7.02 | .73 | .75 | .85 |
| Repl.5 | 48.06 | 7.05 | 51.10 | 8.47 | 49.74 | 7.84 | .70 | .69 | .81 |

Sample Size = 25

| Data Set | Section 1 | | Section 2 | | Section 3 | | Section Corrs. | | |
|---|---|---|---|---|---|---|---|---|---|
| | X | SD | X | SD | X | SD | $r_{12}$ | $r_{23}$ | $r_{13}$ |
| Parms. | 47.61 | 6.83 | 50.23 | 7.73 | 49.55 | 7.27 | .67 | .69 | .81 |
| Repl.1 | 47.20 | 6.53 | 47.92 | 6.36 | 48.88 | 6.05 | .77 | .63 | .79 |
| Repl.2 | 42.20 | 6.86 | 46.76 | 6.95 | 46.56 | 6.93 | .71 | .66 | .85 |
| Repl.3 | 47.04 | 7.98 | 48.88 | 7.09 | 48.20 | 6.68 | .78 | .85 | .81 |
| Repl.4 | 51.40 | 4.74 | 51.96 | 5.65 | 51.48 | 5.72 | .44 | .62 | .71 |
| Repl.5 | 51.00 | 6.62 | 53.32 | 9.15 | 50.72 | 7.84 | .69 | .71 | .81 |

Table A.4
Summary Statistics for the Simulated Data Sets
for Test Center D at Sample Sizes of 100, 50, and 25

Sample Size = 100

| Data Set | Section 1 X | Section 1 SD | Section 2 X | Section 2 SD | Section 3 X | Section 3 SD | Section Corrs. $r_{12}$ | Section Corrs. $r_{13}$ | Section Corrs. $r_{23}$ |
|---|---|---|---|---|---|---|---|---|---|
| Parms. | 47.61 | 6.83 | 50.23 | 7.73 | 49.55 | 7.27 | .67 | .69 | .81 |
| Repl.1 | 46.10 | 7.16 | 49.64 | 8.16 | 49.35 | 7.32 | .64 | .66 | .82 |
| Repl.2 | 46.72 | 6.39 | 50.82 | 7.39 | 50.94 | 6.26 | .61 | .57 | .78 |
| Repl.3 | 45.62 | 6.32 | 50.81 | 8.05 | 50.28 | 7.73 | .70 | .69 | .86 |
| Repl.4 | 43.95 | 7.93 | 50.97 | 9.23 | 49.93 | 8.07 | .75 | .71 | .84 |
| Repl.5 | 42.79 | 7.12 | 50.25 | 7.77 | 49.24 | 7.27 | .64 | .72 | .85 |

Sample Size = 50

| Data Set | Section 1 X | Section 1 SD | Section 2 X | Section 2 SD | Section 3 X | Section 3 SD | Section Corrs. $r_{12}$ | Section Corrs. $r_{23}$ | Section Corrs. $r_{13}$ |
|---|---|---|---|---|---|---|---|---|---|
| Parms. | 47.61 | 6.83 | 50.23 | 7.73 | 49.55 | 7.27 | .67 | .69 | .81 |
| Repl.1 | 46.02 | 6.11 | 49.78 | 7.52 | 49.16 | 5.37 | .65 | .60 | .82 |
| Repl.2 | 45.70 | 7.48 | 50.88 | 6.40 | 50.48 | 6.31 | .73 | .70 | .79 |
| Repl.3 | 45.74 | 6.57 | 50.66 | 8.35 | 49.44 | 8.18 | .66 | .70 | .82 |
| Repl.4 | 44.92 | 6.35 | 51.00 | 8.05 | 50.68 | 6.61 | .69 | .67 | .81 |
| Repl.5 | 43.08 | 7.42 | 49.34 | 7.05 | 49.64 | 6.36 | .79 | .76 | .84 |

Sample Size = 25

| Data Set | Section 1 X | Section 1 SD | Section 2 X | Section 2 SD | Section 3 X | Section 3 SD | Section Corrs. $r_{12}$ | Section Corrs. $r_{23}$ | Section Corrs. $r_{13}$ |
|---|---|---|---|---|---|---|---|---|---|
| Parms. | 47.61 | 6.83 | 50.23 | 7.73 | 49.55 | 7.27 | .67 | .69 | .81 |
| Repl.1 | 44.92 | 6.95 | 48.88 | 6.94 | 47.20 | 6.12 | .68 | .64 | .71 |
| Repl.2 | 45.68 | 6.52 | 50.32 | 9.56 | 50.36 | 7.37 | .64 | .71 | .82 |
| Repl.3 | 45.96 | 7.06 | 51.36 | 7.03 | 50.04 | 5.97 | .66 | .63 | .77 |
| Repl.4 | 44.64 | 7.66 | 50.32 | 8.42 | 51.12 | 8.49 | .88 | .76 | .87 |
| Repl.5 | 43.92 | 6.03 | 51.16 | 7.52 | 49.24 | 7.17 | .70 | .87 | .74 |

Table A.5
Summary Statistics for the Simulated Data Sets
for Test Center E at Sample Sizes of 100, 50, and 25

Sample Size = 100

| Data Set | Section 1 | | Section 2 | | Section 3 | | Section Corrs. | | |
|---|---|---|---|---|---|---|---|---|---|
| | X | SD | X | SD | X | SD | $r_{12}$ | $r_{13}$ | $r_{23}$ |
| Parms. | 47.61 | 6.83 | 50.23 | 7.73 | 49.55 | 7.27 | .52 | .54 | .81 |
| Repl.1 | 45.73 | 7.11 | 48.54 | 8.44 | 48.18 | 7.80 | .56 | .58 | .83 |
| Repl.2 | 47.01 | 5.81 | 50.94 | 7.84 | 49.77 | 6.68 | .55 | .52 | 76 |
| Repl.3 | 44.78 | 7.04 | 50.00 | 7.75 | 49.98 | 6.66 | 59 | .52 | .84 |
| Repl.4 | 43.47 | 5.91 | 49.35 | 7.21 | 49.45 | 6.40 | .32 | .47 | .77 |
| Repl.5 | 41.37 | 6.24 | 50.20 | 7.47 | 50.48 | 6.85 | .54 | .46 | .83 |

Sample Size = 50

| Data Set | Section 1 | | Section 2 | | Section 3 | | Section Corrs. | | |
|---|---|---|---|---|---|---|---|---|---|
| | X | SD | X | SD | X | SD | $r_{12}$ | $r_{23}$ | $r_{13}$ |
| Parms. | 47.61 | 6.83 | 50.23 | 7.73 | 49.55 | 7.27 | .52 | .54 | .81 |
| Repl.1 | 47.10 | 8.31 | 50.02 | 8.59 | 49.42 | 8.85 | .69 | .77 | .88 |
| Repl.2 | 47.28 | 7.51 | 52.30 | 7.98 | 50.94 | 7.29 | .60 | .59 | .74 |
| Repl.3 | 44.48 | 5.93 | 50.54 | 7.96 | 49.58 | 7.16 | .57 | .51 | .77 |
| Repl.4 | 43.20 | 6.85 | 47.98 | 9.16 | 48.12 | 7.60 | .35 | .50 | .88 |
| Repl.5 | 42.48 | 6.51 | 49.56 | 9.77 | 47.96 | 8.80 | .53 | .58 | .86 |

Sample Size = 25

| Data Set | Section 1 | | Section 2 | | Section 3 | | Section Corrs. | | |
|---|---|---|---|---|---|---|---|---|---|
| | X | SD | X | SD | X | SD | $r_{12}$ | $r_{23}$ | $r_{13}$ |
| Parms. | 47.61 | 6.83 | 50.23 | 7.73 | 49.55 | 7.27 | .52 | .54 | .81 |
| Repl.1 | 45.88 | 6.86 | 48.96 | 5.16 | 47.40 | 4.99 | .48 | .59 | .73 |
| Repl.2 | 44.76 | 6.93 | 49.72 | 7.58 | 48.44 | 7.14 | .59 | .53 | .88 |
| Repl.3 | 44.72 | 8.13 | 50.60 | 6.86 | 49.88 | 7.28 | .69 | .55 | .91 |
| Repl.4 | 41.32 | 5.99 | 50.16 | 7.67 | 49.40 | 6.54 | .60 | .61 | .86 |
| Repl.5 | 41.96 | 8.10 | 49.76 | 8.74 | 47.80 | 6.64 | .67 | .50 | .84 |

TOEFL is a program of
Educational Testing Service
Princeton, New Jersey, USA



57906 06415 • D21P5 • 275705 • Printed in U S A